Europäisches Patentamt	European Patent Office	Office européen des brevets

Bescheinigung Certificate Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. Patent application No. Demande de brevet n°

99400219.4

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

I.L.C. HATTEN-HECKMAN

DEN HAAG, DEN
THE HAGUE,
LA HAYE, LE 13/10/99

Europäisches Patentamt
European Patent Office
Office européen des brevets

Blatt 2 der Bescheinigung
Sheet 2 of the certificate
Page 2 de l'attestation

Anmeldung Nr.:
Application no.: 99400219.4
Demande n°:

Anmeldetag:
Date of filing: 01/02/99
Date de dépôt:

Anmelder:
Applicant(s):
Demandeur(s):
Koninklijke Philips Electronics N.V.
5621 BA Eindhoven
NETHERLANDS

Bezeichnung der Erfindung:
Title of the invention:
Titre de l'invention:
Mpeg-7 descriptor for camera motion

In Anspruch genommene Priorität(en) / Priority(ies) claimed / Priorité(s) revendiquée(s)

Staat:	Tag:	Aktenzeichen:
State:	Date:	File no.
Pays:	Date:	Numéro de dépôt:

Internationale Patentklassifikation:
International Patent classification:
Classification internationale des brevets:
/

Am Anmeldetag benannte Vertragstaaten:
Contracting states designated at date of filing: AT/BE/CH/CY/DE/DK/ES/FI/FR/GB/GR/IE/IT/LI/LU/MC/NL/PT/SE
Etats contractants désignés lors du depôt:

Bemerkungen:
Remarks:
Remarques:

Descriptor for Camera Motion

MPEG-7 Evaluation Ad Hoc meeting

p635

Benoit MORY, Laboratoires d'Electronique Philips

This is a draft version of the final document.

TABLE OF CONTENT

1. INTRODUCTION

1.1. Feature used, nature of proposal

This proposal aims at submitting to the MPEG-7 group a representation of the camera motion within any sequence of frames in a video scene. This proposal is not easy to classify as a descriptor or a description scheme. As a consequence, we will study in this paper how the global motion representation that we propose can cope with both the descriptors and description schemes requirements. Moreover, we will also propose a way to describe the motion of the camera within a shot, which is the most commonly used partitioning of the video. This latter description will be a description scheme using the basic camera motion descriptor.

1.2. Feature relevance : Importance of camera motion

Camera operations are very important from a video indexing viewpoint. Since Objects motion and global motion are the key features that make the difference between still images and video, any indexing system based on the video content should include a way to efficiently represent motion in a wide sense. As far as the motion of the camera is concerned, it is obvious that the parts of the video in which camera is static and those in which the camera is travelling or panning don't share the same sense in terms of spatio-temporal content. Like any other discriminating feature, the global motion must be described and represented in the future MPEG-7 framework.

1.3. General scope

The scope of this submission is very large, as it can address any type of video and any type of application in which the motion of the camera can be an issue. In video archives, adding a description of the global motion allows the users, either non-expert or professional, to perform queries that take into account the motion of the camera. Those queries, mixed with the description of other features, should permit to retrieve video shots according to information directly or semantically related to the camera motion. For example, a close-up of an object is likely to be preceded by a zoom in and a shot of a landscape generally involves a pan. [TO BE CONTINUED]

1.4. Outline of the proposal

In this submission, a descriptor for the representation of the camera motion within video scenes is proposed. In section 2, the MPEg-7 requirements and evaluation criteria that are relevant to this proposal are listed and studied with respect to our description. In section 3.1, the representation itself is developed and our technical choices are explained. In section 3.2, a brief overview of the existing methods for extracting camera parameters automatically is given. In section 3.3, an original matching method based on the query parameters is described. [TO BE CONTINUED]

2. WHAT IS SATISFIED

In this section, all the requirements and evaluation criteria (taken from the official MPEG-7 documents ISO/IEC JTC1/SC29/WG11/N2461 and ISO/IEC JTC1/SC29/WG11/N2463) that are relevant to this proposal are listed and, if possible, are shown to be satisfied by the camera motion description proposed. Both description schemes and descriptors requirements are taken into account, as our proposal can be placed somewhere in between.

We try here to study, for each requirement and evaluation criteria related to our proposal, if it is fulfilled and how it offers a good solution for the given issue. The requirements that are not relevant are omitted.

2.1. Requirements

The requirements potentially relevant to our proposal are the ones contained in the following sections of the MPEG-7 Requirements document ISO/IEC JTC1/SC29/WG11/N2461:

- Visual Requirements,
- Descriptors and Description Schemes General Requirements,
- Descriptors and Description Schemes Functional Requirements,

2.1.1. Visual Requirements

Type of features

"MPEG-7 shall at least support visual descriptions allowing the following features (mainly related to the type of information used in the queries): [...] Motion (for retrievals using temporal composition information.) [...]"

The feature targeted by our proposal, namely the camera motion, is obviously related to motion

Data visualisation using the description

"MPEG-7 shall support a range of multimedia data descriptions with increasing capabilities in terms of visualisation. This means that MPEG-7 data descriptions shall allows a more or less sketchy visualisation of the indexed data."

As far as visualisation is concerned, one can imagine to textually or graphically represent the camera operations described to obtain a kind of summary of the global motion of the video, for instance inside a story board.

Visual data formats

"MPEG-7 shall support the description of the following visual data formats: digital video and film, such as MPEG-1, MPEG-2 or MPEG-4; analogue video and film, still pictures in electronic such as JPEG, paper or other format; graphics, such as CAD, 3D models, notably VRML; composition data associated to video; others to be defined."

The description submitted here targets all the video data formats, digital as well as analogue, since it is related to the video content itself. However, the automatic extraction can be easier on digital compressed video data, in which motion information is already included in the content (e.g. motion vectors of MPEG-1, 2 and 4 format).

Visual data classes

"MPEG-7 shall support descriptions specifically applicable to the following classes of visual data: natural video, still pictures, graphics, animation (2-D), three-dimensional models, composition information."

Our proposal makes sense being applied on any animated visual data, like natural video, animations or cartoons.

2.1.2. General Requirements (D & DS)

Types of features

The feature described by our proposal, following the MPEG-7 requirements feature types classification, is a N-dimensional spatio-temporal feature.

Abstraction levels for Multimedia material

"[...] describe material hierarchically according to abstraction levels of information to efficiently represent user's information need at different levels."

The description we propose is generic and can be used inside a more or less hierarchical scheme. It can represent the camera motion in a wide range of temporal granularity. Typically, the same descriptor can represent the global motion types and speeds of a film, a video shot, a micro-segment within a shot or even a single frame. As a consequence, this proposal offers different abstraction levels for the targeted feature.

Cross-modality

"[...] audio, visual, or other descriptors which allow queries based on visual descriptions to retrieve audio data and vice versa."

Queries based on camera motion, as mentioned in Section 1.3, can allow the retrieval of completely different particular features of the visual content. For example, some assertions like *"a close-up of an object is likely to be preceded by a zoom"* and "*a shot of a landscape generally involves a pan*" lead us to oversee the use of camera motion operations description to help or refine searches in which different types of features are involved.

Feature priorities

"[...] prioritisation of features in order that queries may be processed more efficiently. The priorities may denote some sort of level of confidence, reliability, etc"

The prioritisation of features is not directly inside the scope of the proposal. However, the descriptor structure involves more or less a distribution of the different camera motion parameters (see section 3.1, MotionTypesHistogram), which can be interpreted as a kind of prioritisation of these parameters from the retrieval viewpoint.

Feature hierarchy

"[...] hierarchical representation of different features in order that queries may be processed more efficiently in successive levels where N level features complement (N-1) level features."

The camera motion description itself is not designed following a hierarchical scheme. However, in an higher temporal level of description, one can imagine to represent for example the motion of a video scene, inside which each video shot is also described, and so forth recursively until the frame level. Such a hierarchical scheme could allow a more efficient processing of the data in terms of query.

Descriptor scalability

"[...] scalable Ds in order that queries may be processed more efficiently in successive layers where N-layer description data is an enhancement/refinement of (N-1) layer description data. An example is MPEG-4 shape scalability."

The above reasoning is also relevant for the scalability.

Description Schemes with multiple levels of Abstraction.

"[...] DSs which provide abstractions at multiple levels for instance a coarse-to-fine description [...]"
see ***Abstraction levels for Multimedia material***

Description of temporal range

"[...] association of Ds to different temporal ranges, both hierarchically (Ds are associated to the whole data or a temporal sub-set of it) as well as sequentially (Ds are successively associated to successive time periods)."

This requirement is completely fulfilled by the current proposal. Indeed, the camera motion descriptor can be associated to different temporal ranges of the video material, from the whole video (e.g. this particular film has been shot using always a fixed camera) to the frame level (very fine description). More over, the description can also be associated to successive time periods, like different micro-clusters within a shot (e.g. this shot begins with a long zoom of 20 seconds and ends with a short tilt of 2 seconds). The latter description is the target of the so-called "Shot level camera motion DS" presented in section 3.1.2.

Direct data manipulation

"[...] Ds which can act as handles referring directly to the data, [...]"
The proposed descriptor effectively allows the direct data manipulation. As a matter of fact, the time instants defining the temporal window to which the camera motion being described is related is given inside the descriptor. This is precisely the reason why the proposal is somewhere in between a descriptor and a description scheme, but it is also a guarantee of flexibility and genericity.

2.1.3. Functional Requirements (D & DS)

Content-based retrieval

"[. .] effective ('you get what you are looking for and not other stuff') and efficient ('you get what you are looking for, quickly') retrieval of multimedia data based on their contents whatever the semantic involved "

One the main goals of our proposal is precisely to allow the effective and efficient retrieval of the video data based on the camera motion. The effectiveness is mainly guaranteed by the preciseness of the description that takes into account independently all the possible camera motion operations as well as the speeds involved. The efficiency is fairly dependent of the database engine used and the retrieval strategy chosen and is out of the scope of this proposal. However, some results will be shown during the demo in Lancaster.

Similarity-base retrieval

"[...] descriptions allowing to rank-order database content by the degree of similarity with the query."

Similarity-base retrieval is also targeted by our camera motion description. Since a similarity function is being associated to the description, this retrieval as well as the ranking should be possible.

Streamed and stored descriptions

No particular aspect of our proposal seems to forbid both the storage and the streaming of the description.

Referencing analog data

"[...] reference, describe AV objects and time references of analog format."
Once again, there is no limitation for referencing analogue data with the description proposed.

Linking

"[...] source data located using MPEG-7. [...] mechanism to link to related information."
This descriptor proposal allows the precise locating of the referenced data, since the time instants defining the temporal window during which the description is valid is included in the description.

2.1.4. Coding Requirements (D & DS)

Description efficient representation

"[...] efficient representation of data descriptions."

The coding of the description is not covered by this proposal, but should not be a critical issue since only a limited number of simple data types are involved.

Description extraction

"[...] Ds and DSs easily extractable from uncompressed and compressed data, according to several widely used formats."

The automatic camera motion parameters extraction has been widely studied ([GIVE REFERENCES]) and the proposed description is potentially extractable from any video format, compressed or not.
However, the automatic extraction can be easier and particularly more efficient on digital compressed video data, since motion information is already included in the content (e.g motion vectors of MPEG-1, 2 and 4 format).

2.2. Evaluation Criteria

2.2.1. Evaluation Criteria for Descriptors

Feature relevance

The feature captures important characteristic(s) of the AV material.

The feature relevance according to the MPEG-7 video indexing framework has been explained in section 1.2.

Effectiveness

Gives better retrieval accuracy (e.g. precision, recall rate) with respect to other descriptors for the same feature.

The effectiveness of the descriptor proposed is being studied and should give good results, since each motion component is described independently and precisely. However, the retrieval accuracy will

certainly depend on the searching strategy involved in the database system and results will be shown during the demo in Lancaster.

Application domain

The Descriptor is applicable to a wide range of application domains.

The application domain is very large since the camera motion is a key feature for all the video content-based application, from query/retrieval systems to video surveillance or video edition.

Expression efficiency

The Descriptor expresses the given feature(s) precisely, and completely.

The efficiency of the camera motion description developed in this submission is mainly due to the fact that all the possible motion parameters as well as the speeds involved are independently described. The precision on the movements speed is half a pixel per frame, which should be sufficient in all the possible applications. The temporal precision depends on the time granularity chosen and can achieve the frame level (One camera motion description for each frame).

Processing efficiency

- ✓ *An efficient Descriptor value calculation method exists.*
- ✓ *An efficient matching method (allowing rank ordering) associated with this Descriptor exists.*

Several camera motion parameters calculation exist ([GIVE REF.]) and, for most of them, appear to be efficient. However, separating all the distinct camera motion types is a rather difficult task that is being studied and is out of the strict scope of this proposal.

The simplicity of the descriptor and its comprehensiveness makes the matching quite straightforward and allows the query to be parameterized, in order to satisfy a wide amount of possible queries.

Scalability

- ✓ *For a given application, the performance does not degrade with larger amount of data*

The scalability in terms of amount of data is not in the scope of this proposal.

Multi-level representation

The Descriptor represents the features at multiple levels of abstraction.

The description proposed in this paper offers the possibility to be used inside a more or less hierarchical scheme. Indeed, It can represent the camera motion in a wide range of temporal granularity. Typically, the same descriptor can represent the global motion types and speeds of a film video (e.g. this particular film has been shot using always a fixed camera), a video shot, a micro-segment within a shot (e g. this shot begins with a long zoom of 20 seconds and ends with a short tilt of 2 seconds) or even a single frame (very fine description for particular application). As a consequence, this proposal offers different abstraction levels for the targeted feature.

2.2.2. Evaluation Criteria for Description Schemes

Since some evaluation criteria are common to both descriptors and description schemes, references to the latter section have sometimes been used.

Effectiveness of the DS in accomplishing its stated purpose.

See the above section (2.2.1)

Application domain

The DS is applicable for a wide range of applications. "Applicable" means directly usable or usable as a component of a larger DS.

See the above section (2.2.1)

Comprehensiveness

The DS provides an off the shelf solution for a given application domain. For this application domain, it takes into account relevant Descriptors and relevant relations between the Descriptors.

The comprehensiveness of the description proposed (camera motion Descriptor and shot-level camera motion Description Scheme) is guaranteed by an exhaustive and independent description of all the possible camera motion operations. The comprehensiveness of the Description Scheme representing the different camera movements within a shot is fairly related to the consistence of the temporal partitioning involved (The micro-clusters must be chosen in order to keep a certain consistency in terms of camera motion) but is made possible by the structure used itself.

Abstraction at Multiple Hierarchical Levels

The DS can provide abstractions at multiple levels. An example is a hierarchical scheme where the base layer gives a coarse description and successive layers give more refined descriptions. The type of hierarchy used is appropriate for the purpose of the DS. Descriptors within the DS are amenable to being prioritized.

See the above section (Multi-level representation)

Flexibility

Part of the DS can be used effectively:

- ✓ *Ability to instantiate a part of a DS.*
- ✓ *Ability to efficiently access a part of a DS.*
- ✓ *Ability to accept additional Descriptors; existing Descriptors can be replaced with new Descriptors.*

This kind of flexibility is not relevant for the description proposed (it aims at describing one single feature and makes sense in its globality).

Extensibility

The DS is easily extensible to other DSs (in a way similar to inheritance in Object-Oriented Programming).

The extensibility of the camera motion description is shown by the example of a shot-level camera motion Description Scheme given in section 3.1.2. . Indeed, this description scheme is an extension of the basic descriptor we propose. Obviously, it can also be included in other description schemes as one out of many relevant features of the video data.

Scalability

- ✓ *For a given application, the performance does not degrade with larger amount of data.*
- ✓ *Scalability across different applications (down or up).*

This kind of scalability is not relevant to our proposal.

Simplicity

A minimal number of Descriptors and possible relationships are used to meet the needs of a particular application domain.

[TO BE DONE]

3. TECHNOLOGY DESCRIPTION

3.1. Representation

3.1.1. Camera operations

Regular camera operations include panning, zooming, tracking and dollying together with the numerous possible combinations of these operations. The eight well-known basic camera operations (see figure 1.1, 1.2 and 1.3) generally defined are fixed, panning (horizontal rotation), tracking (horizontal transverse movement, also called travelling in the film language), tilting (vertical rotation), booming (vertical transverse movement), zooming (changes of the focal length), dollying (translation along the optical axis) and rolling (rotation around the optical axis).

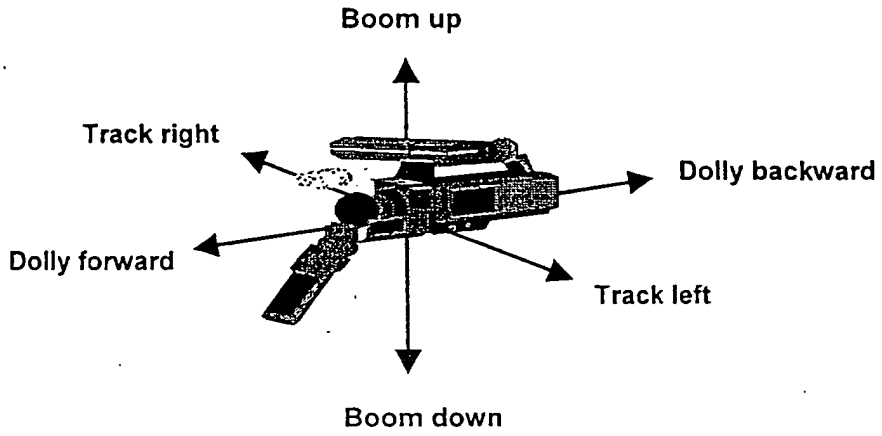


Figure 1.1

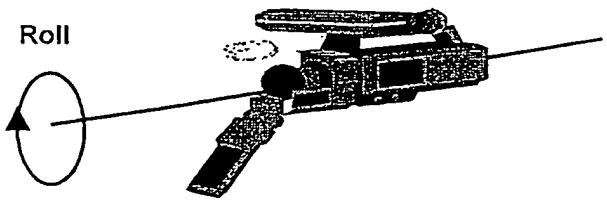


Figure 1.2

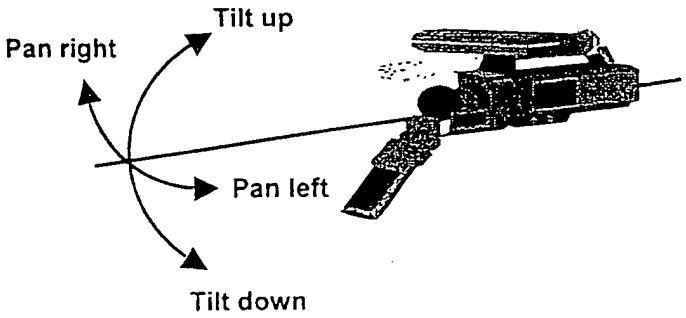


Figure 1.3

Fixed operation is very common and doesn't need further explanation. Panning and tilting are often used, particularly when the camera center is fixed (on a tripod for instance), and allow the following of an object or the view of a large scene (a landscape or a skyscraper for instance). Zooming are often used to focus the attention on a particular part of a scene. Tracking and dollying are most of the times used to follow moving objects (e.g travelling). Rolling is for instance the result of an acrobatic sequence shot from an airplane All seven camera motion operations (fixed is straightforward) lead to different patterns of the motion vectors field in the image that can be automatically extracted (see section 3.2)

3.1.2. Camera motion descriptor

Given the regular camera operations described in the latter section, a generic descriptor for camera motion should be able to represent all those motion types independently, in order to handle every combination of them without any restriction. The scheme we propose is compliant with this approach. Each motion type, except fixed camera, is oriented and can be subdivided into two components that stand for two different directions. Indeed, panning and tracking can be either left or right, tilting and booming can be either up or down, zooming can be in or out, dollying can be forward or backward and rolling can be either left (direct sense) or right (reverse sense) (see. Fig 1.1, 1.2 and 1.3). The distinction between the 2 possible directions allows us to use always positive values for the 13 motion types and to represent them in a way similar to an histogram (see MotionTypesHistogram). Here under a data type for representing the different movements (MotionTypesHistogram) is given in the UML language. This data type will be used inside our basic camera motion descriptor.

MotionTypesHistogram
PAN_LEFT : unsigned short.
PAN_RIGHT : unsigned short.
TRACK_LEFT : unsigned short.
TRACK_RIGHT : unsigned short.
TILT_DOWN : unsigned short.
TILT_UP : unsigned short.
BOOM_DOWN : unsigned short.
BOOM_UP : unsigned short.
ZOOM_IN : unsigned short.
ZOOM_OUT : unsigned short.
ROLL_LEFT : unsigned short.
ROLL_RIGHT : unsigned short.
DOLLY_FORW : unsigned short.
DOLLY_BACK : unsigned short.
FIXED : unsigned short.

Instantaneous motion

Each motion type is supposed to be independent and to have its own speed, that we aim at describing in an unified way. As the local speed induced by each motion type can depend on the scene depth (in the case of translations) or on the image point location (in the case of zooming, dollying and rotations), we have chosen a common unit to represent it. A speed will be represented by a pixel/frame value in the image plane, which is close to our speed perception. In the case of translations, the motion vectors magnitude is to be averaged on the whole image, because the local speed depends on the objects depth [L.H. '80]. In case of rotations like panning or tilting, the speed will be the one induced at the centre point of the image, where there is no distortion due to side effects. In case of zooming, dollying or rolling, the motion vectors field is divergent (more or less proportional to the distance to the image centre) We'll choose then to represent the speed by the pixel displacement of the image corners.

Each motion type speed is then represented by a pixel-displacement value. We propose, so as to meet the efficiency requirements, to work at the half-pixel accuracy. As a consequence, in order to work with integer values, speeds will always be rounded to the closest half-pixel value and multiplied by 2

Given these definitions, any instantaneous camera motion can be represented by a histogram of the motion types (see UML description above) in which the values correspond to half-pixel displacements. It is obvious that the FIXED field makes no sense in terms of speed. This is the reason why a specific data type is required, in which FIXED is removed (see here under, CameraMotionSpeed)

CameraMotionSpeed	
PAN_LEFT :	unsigned short.
PAN_RIGHT :	unsigned short.
TRACK_LEFT :	unsigned short.
TRACK_RIGHT :	unsigned short.
TILT_DOWN :	unsigned short.
TILT_UP :	unsigned short.
BOOM_DOWN :	unsigned short.
BOOM_UP :	unsigned short.
ZOOM_IN :	unsigned short.
ZOOM_OUT :	unsigned short.
ROLL_LEFT :	unsigned short.
ROLL_RIGHT :	unsigned short.
DOLLY_FORW :	unsigned short.
DOLLY_BACK :	unsigned short.

Long-term representation

Working only with descriptions of instantaneous movements would be very heavy and time-consuming. We aim in this proposal at defining a description more or less hierarchical, that is to say handling the representation of the camera motion at any temporal granularity.

Given a temporal window of the video data $[n_0, n_0 + N]$ (N is the total number of frames of the window), we suppose that we know the speeds of each motion type for each frame. We can then compute the number of frames N_{motion_type} in which each motion type has a significant speed and represent the temporal presence by a percentage, defined as follows (e.g. for a panning movement) :

$$T_{panning} = \frac{N_{panning}}{N}$$

The temporal presence of all the possible camera motions would then be represented by a MotionTypesHistogram in which the values, between 0 and 100, correspond to a percentage. Obviously, if the window is reduced to a single frame, the values can only be 0 or 100, depending on the fact that the given movement is present or not in the frame.

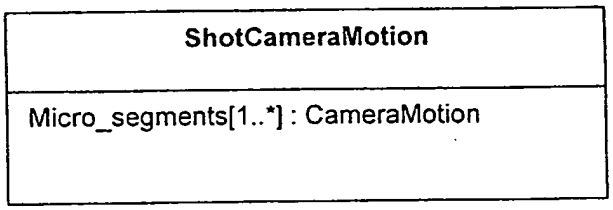
Finally, in order to directly access the represented video data and to allow efficient comparisons between descriptors, we add to the description the temporal boundaries that define the window being described, which can be either a entire video, a shot, a micro-segment (part of a shot) or a single frame. The UML description of the so-defined camera descriptor is given here under. Obviously, the speeds correspond to the instantaneous speeds averaged on the whole temporal window (when the given motion type is present). Note that there is no particular constraint on the TimeStamp format, which can be any of the ones that should certainly be addressed by other proposals.

CameraMotion	
start_time :	TimeStamp
stop_time :	TimeStamp
temporal_presence :	MotionTypesHistogram
speeds :	CameraMotionSpeed

3.1.3. Shot-level camera motion : A Description Scheme

In the following section , we propose a simple Description Scheme that uses the camera motion descriptor we just defined. This Description Scheme aims at representing the motion of the camera of a widely used temporal partitioning of the video, namely the shot. A shot, in the film field, is a sequence of frames in which there is no discontinuity, and thus offers a natural index when dividing a video into coherent temporal elements.

It is straightforward that a shot, like any other temporal window of a video, can be represented by the camera motion descriptor itself, but with the risk to mix many motion types together. Imagine a shot beginning with a long zoom in followed by a fixed camera and ending with a combination of tilting and panning operation. The descriptor that we described in section 3.1.3. will mix all the temporal information and will just be able to give you an idea of each motion relative importance within the shot, without any knowledge about the exact succession of the movements. Such a shot would really benefit from its subdivision into several parts. In this case, 3 micro-clusters (shot subdivisions) should be defined, and the camera motion should be described for each of them. This is the reason why we propose to represent the global motion at the shot level by a set (an array) of several micro-segments corresponding to different descriptions. Ideally, the motion of the camera should be consistent in each of the micro-clusters, but any subdivision can be performed. The UML description of such a Description Scheme is presented here under.



[GIVE HERE THE REPRESENTATION OF THE EXAMPLE]

3.2. Extraction

The extraction of the camera motion parameters has been studied, for instance in [Akutsu '92], [L.H. '80] [Srinivasan '97], [Bouthemy], [Idris '97]. [TO BE DEVELOPED]

3.3. Matching

3.3.1. Comparing two camera motion Ds

In this section, we give an example of similarity function that could be associated we the descriptor defined in the paper. As the sense of similarity fairly depends on the type of query being performed, we firstly identify some query criteria so as be able to integrate some query parameters in the similarity function.

3 different types of query parameters have been included :

for each motion type, the presence can be taken into account or not :

$\forall n / 1 \le n \le 13,$

$\alpha_n = 1$ if the motion type is to be taken into account

$\alpha_n = 0$ otherwise

For each motion type, the speed can be discriminating or not :

$\forall n / 1 \le n \le 12,$

$\beta_n = 1$ if the speed of the motion type is to be taken into account

$\beta_n = 0$ otherwise

Note that $\alpha_n = 0 \Rightarrow \beta_n = 0$

The whole duration of the temporal segment being described can have an importance or not.

$\gamma = 1$ if the whole duration is to be taken into account

$\gamma = 0$ otherwise

Let $CM1$ and $CM2$ be two descriptors of two distinct temporal video segments.

Let T_i^{CM1}, S_i^{CM1} (respectively T_i^{CM2}, S_i^{CM2}) be the Temporal presence factors and the Speeds of each motion type.

Let D^{CM1} (respectively D^{CM2}) be the whole duration of the considered temporal segment.

The similarity $L_1(CM1, CM2, \alpha, \beta, \gamma)$ between the 2 descriptors can be computed as follows :

$$L_1(CM1, CM2, \alpha, \beta, \gamma) = \frac{1}{2+\gamma}\left[\frac{\sum_{i=1}^{13}\alpha_i\left(1-\frac{\left|T_i^{CM1}-T_i^{CM2}\right|}{T_i^{CM1}+T_i^{CM2}}\right)}{\sum_{i=1}^{13}\alpha_i} + \frac{\sum_{i=1}^{12}\beta_i\left(1-\frac{\left|S_i^{CM1}-S_i^{CM2}\right|}{S_i^{CM1}+S_i^{CM2}}\right)}{\sum_{i=1}^{13}\beta_i} + \gamma\left(\frac{\left|D^{CM1}-D^{CM2}\right|}{D^{CM1}+D^{CM2}}\right)\right]$$

In the above equation, we have assumed that the whole duration is not null and that at least one query parameter α_i (respectively β_i) out of 13 (respectively out of 12) is not null. If it is the case, the similarity function takes obviously a more simple form.

Note that the so-defined similarity function is between 0 (complete mismatch) and 1 (complete matching). It allows a similarity-based retrieval based on camera motion, as well as the rank-ordering of the query results. Many types of queries are foreseen, made possible by the 3 types of query parameters that give different meanings to the similarity.

In this section, we try to define a similarity function for the comparison between 2 shot-level camera motion Description Schemes that have been defined in section 3.1.3.

Let $C^{(1)}$ and $C^{(2)}$ be the two considered Description Schemes.

Let $C_i^{(1)}, i \in \{1,2,..N_1\}$ and $C_j^{(2)}, j \in \{1,2,..N_2\}$ be the 2 sets of CameraMotion descriptors.

One out of the 2 Description Schemes must be associated to a set query parameters, already defined in the latter section, for each temporal micro-segment (for each individual descriptor). Practically, this will be the Query descriptor, or the given example in the case of query by example. Let us choose $C_1^{(1)}$.

Let $\alpha_{i,m}$, $i \in \{1,2,..N_1\}$, $m \in \{1,2,..13\}$ be the importance given to the presence of each motion type m within each micro-segment i.

Let $\beta_{i,m}$, $i \in \{1,2,..N_1\}$, $m \in \{1,2,..12\}$ be the importance given to the speed of each motion type m within each micro-segment i.

Let β_i , $i \in \{1,2,..N_1\}$ be the importance given to each micro-segment's duration.

The similarity function can be defined as follows :

$$L_{DS}\left(C^{(1)}, C^{(2)}, \alpha, \beta, \gamma\right) = \frac{1}{N_1} \sum_{i=1}^{N_1} \max_j \left\{ L_{ij}\left(C_i^{(1)}, C_j^{(2)}, \alpha_i, \beta_i, \gamma_i\right) \right\}$$

This similarity function is in between 0 (complete mismatch) and 1 (complete matching) and should allow similarity-based retrieval as well as rank-ordering according to the camera motion of a video shot. However, this function does not take into account the temporal succession of the different micro-segments within the shot, and should then be improved (a shot with successively a zoom and a pan will be considered as equal to the one with successively a pan and a zoom).

4. CONTENT SET USED FOR EVALUATION

5. OTHER REMARKS

6. CONCLUSIONS

7. REFERENCES

[Akutsu '92] A.Akutsu, Y. Tonomura, H.Hashimoto, Y.Ohba, *"Video indexing using motion vectors"* SPIE Vol. 1818 Visual Communications and Image Processing '92, pp.1522-1530

[L.H. '80] H.C. Longuet-Higgins, F.R.S., K. Prazdny, *"The interpretation of a moving retinal image"* proc. R. Soc. Lond. Vol. 208 B (July 1980), pp.385-397

[Srinivasan '97] M.V. Srinivasan, S. Venkatesh, R. Hosie, *"Qualitative estimation of camera motion parameters from video sequences"*, Pattern Recognition, Vol. 30, No 4, 1997, pp. 593-606

[Bouthemy] P. Bouthemy, M. Geldon and F. Ganansia, *"A unified approach to shot change detection and camera motion characterization"* IRISA, Publication interne n° 1148

[Idris '97] F. Idris, S. Panchanathan, "Review of Image and Video Indexing Techniques" Journal of visual communications and image representations, Vol. 8, No 2, June 1997, pp.146-166

CLAIMS :

1. A descriptor for the representation of a camera motion within any sequence of frames in a video scene, said descriptor giving a way to describe the motion of the camera within a shot, which is the most commonly used partitioning of the video.

2. A description scheme using said basic camera motion descriptor.

3. For regular camera operation including particularly panning (horizontal rotation), tracking (horizontal transverse movement, also called travellling in the film language), tilting (vertical rotation), booming (vertical transverse movement), zooming (changes of the focal length), dollying (translation along the optical axis) and rolling (rotation around the optical axis), and the numerous possible combinations of these operations, a generic descriptor able to represent all those motion types independently, in order to handle every combination of them without any restriction, the proposed scheme description being characterized in that :

- each motion type, except fixed camera, is oriented and can be subdivided into two components that stand for two different directions : panning and tracking can be either left or right, tilting and booming can be either up or down, zooming can be in or out, dollying can be forward or backward and rolling can be either left (direct sense) or right (reverse sense) ;
- the distinction between the 2 possible directions allows to use always positive values for the 13 motion types and to represent them in a way similar to an histogram ;
- each motion type is assumed to be independent and to have its own speed described in an unified way by choosing a common unit to represent it : a speed will be represented by a pixel/frame value in the image plane (in case of translations, the motion vectors magnitude is to be averaged on the whole image, because the local speed depends on the objects depth ; in case of rotations like panning or tilting, the speed will be the one induced at the centre point of the image, where there is no distortion due to side effect ; in case of zooming, dollying or rolling, the motion vectors field is divergent and more or less proportional to the distance to the image centre, and the speed is represented by the pixel displacement of the image corners) ;
- each motion type speed is represented by a pixel-displacement value working at the half-pixel accuracy : as a consequence, in order to work with integer values, speeds will always be rounded to the closest half-pixel value and multiplied by 2.

4. A descriptor according to anyone of claims 1 and 3, characterized in that the description is hierarchical, that is to say the representation of the camera motion is handled at any temporal granularity.

5. A descriptor according to claim 4, characterized in that, given a temporal window of the video data $[n_0, n_0 + N]$ (N is the total number of frames of the window) and the speeds of each motion type for each frame, the number of frames N_{motion_type} in

CLAIMS :

1. A descriptor for the representation of a camera motion within any sequence of frames in a video scene, said descriptor giving a way to describe the motion of the camera within a shot, which is the most commonly used partitioning of the video.

2. A description scheme using said basic camera motion descriptor.

3. For regular camera operation including particularly panning (horizontal rotation), tracking (horizontal transverse movement, also called travellling in the film language), tilting (vertical rotation), booming (vertical transverse movement), zooming (changes of the focal length), dollying (translation along the optical axis) and rolling (rotation around the optical axis), and the numerous possible combinations of these operations, a generic descriptor able to represent all those motion types independently, in order to handle every combination of them without any restriction, the proposed scheme description being characterized in that :

- each motion type, except fixed camera, is oriented and can be subdivided into two components that stand for two different directions : panning and tracking can be either left or right, tilting and booming can be either up or down, zooming can be in or out, dollying can be forward or backward and rolling can be either left (direct sense) or right (reverse sense) ;
- the distinction between the 2 possible directions allows to use always positive values for the 13 motion types and to represent them in a way similar to an histogram ;
- each motion type is assumed to be independent and to have its own speed described in an unified way by choosing a common unit to represent it : a speed will be represented by a pixel/frame value in the image plane (in case of translations, the motion vectors magnitude is to be averaged on the whole image, because the local speed depends on the objects depth ; in case of rotations like panning or tilting, the speed will be the one induced at the centre point of the image, where there is no distortion due to side effect ; in case of zooming, dollying or rolling, the motion vectors field is divergent and more or less proportional to the distance to the image centre, and the speed is represented by the pixel displacement of the image corners) ;
- each motion type speed is represented by a pixel-displacement value working at the half-pixel accuracy : as a consequence, in order to work with integer values, speeds will always be rounded to the closest half-pixel value and multiplied by 2.

4. A descriptor according to anyone of claims 1 and 3, characterized in that the description is hierarchical, that is to say the representation of the camera motion is handled at any temporal granularity.

5. A descriptor according to claim 4, characterized in that, given a temporal window of the video data $[n_0, n_0 + N]$ (N is the total number of frames of the window) and the speeds of each motion type for each frame, the number of frames N_{motion_type} in

which each motion type has a significant speed is computed and the temporal presence is represented by a percentage, defined as follows (e.g for a panning movement) :

$$T_{panning} = \frac{N_{panning}}{N}$$

the temporal presence of all the possible camera motions being then represented by a MotionTypesHistogram in which the values, between 0 and 100, correspond to a percentage (if the window is reduced to a single frame, the values can only be 0 or 100, depending on the fact that the given movement is present or not in the frame).